文献资源发现服务系统实现探讨

刘敏健 王星
（中国科学技术信息研究所 北京 100038）

摘要：文章简要介绍文献资源发现服务系统的基本概念， 并根据文献资源发现系统的基本原理详细介绍了文献资源发现系统的技术框架，并对其中使用的关键技术进行了简要说明。
关键词：资源发现，数字图书馆，技术体系，关键技术

引言

在数字图书馆环境下，用户为了检索想要的资源，面临着许多困境。图书馆传统的检索系统对于普通用户来说比较复杂，难以快速掌握；用户对图书馆各种文献的元数据标准不清楚，对它们的特点难以区分；用户很难快速定位获取到所需文献的全文等。用户更偏好于像谷歌和百度那样的简单的检索入口。一些数据库商为了解决这些问题，提出了联邦检索解决方案。这种解决方案将一个检索请求以合适的语法进行转换后发送到一组独立的数据库中，合并检索到的检索结果，以简洁统一的格式和最小的重复显示结果，提供一个自动或者用户选择的排序方式对结果集进行排序。[1] 但是这种解决方案也有很多缺点，最主要的缺点是各个数据库的相应速度不一样, 根据木桶定律，系统响应时间由响应速度最慢的检索服务器决定；在分布式的网络条件下，某个检索服务器的宕机可能导致整个检索的失败，这一特性导致整个系统可靠性比较脆弱；除此意外，联邦检索系统还有资源集成没有统一的元数据标准、查询结果去重困难和相关度排序并不是完全相关度排序等问题。[2]
由于以上问题，联邦检索的用户体验并不好。作为新一代的图书馆检索系统，资源发现系统诞生了。它是一种深度整合图书馆各种类型资源、提供单一入口的学术资源发现服务平台，它能帮助读者快捷、准确地在海量信息资源中查找所需文献，提供最合适的获取服务集成，并在查找过程中获得最佳体验。为了适应网络时代用户的检索习惯，资源发现系统提供了类似 Google 的简单检索框，极大简化了相对比较复杂的图书馆传统检索界面。目前国际上比较著名的资源发现系统有 Summon、Primo、EDS 等。

1、 文献资源发现服务系统的特点

1. 必须提供单一的检索接口

在当前的互联网时代，用户已经习惯了提供单一检索入口的搜索引擎。提供类似 Google 或者百度一样的单一输入框检索引擎，有助于提高用户的使用体验。用户可以简单的输入一个检索词，检索出各种类型的文献，在其中挑选自己需要的文献。

2. 能够帮助用户快捷、准确找到所需文献元数据

用户需要在海量文献数据种快速定位到自己所需要的文献。这需要发现系统检索速度要快；返回结果要全面准确；结果的排序规则要合理。

3. 能够帮助用户方便的获取到全文文件

用户使用资源发现服务系统的最终目的是获取所需要的全文，所以系统要提供最合适

的获取全文服务集成，并使得用户在获取过程中获得最佳体验。

2、 资源发现系统的技术框架

文献资源发现服务系统的技术框架大致可以按照表示层、应用层、数据层的三层结构划分。如图 1 所示：

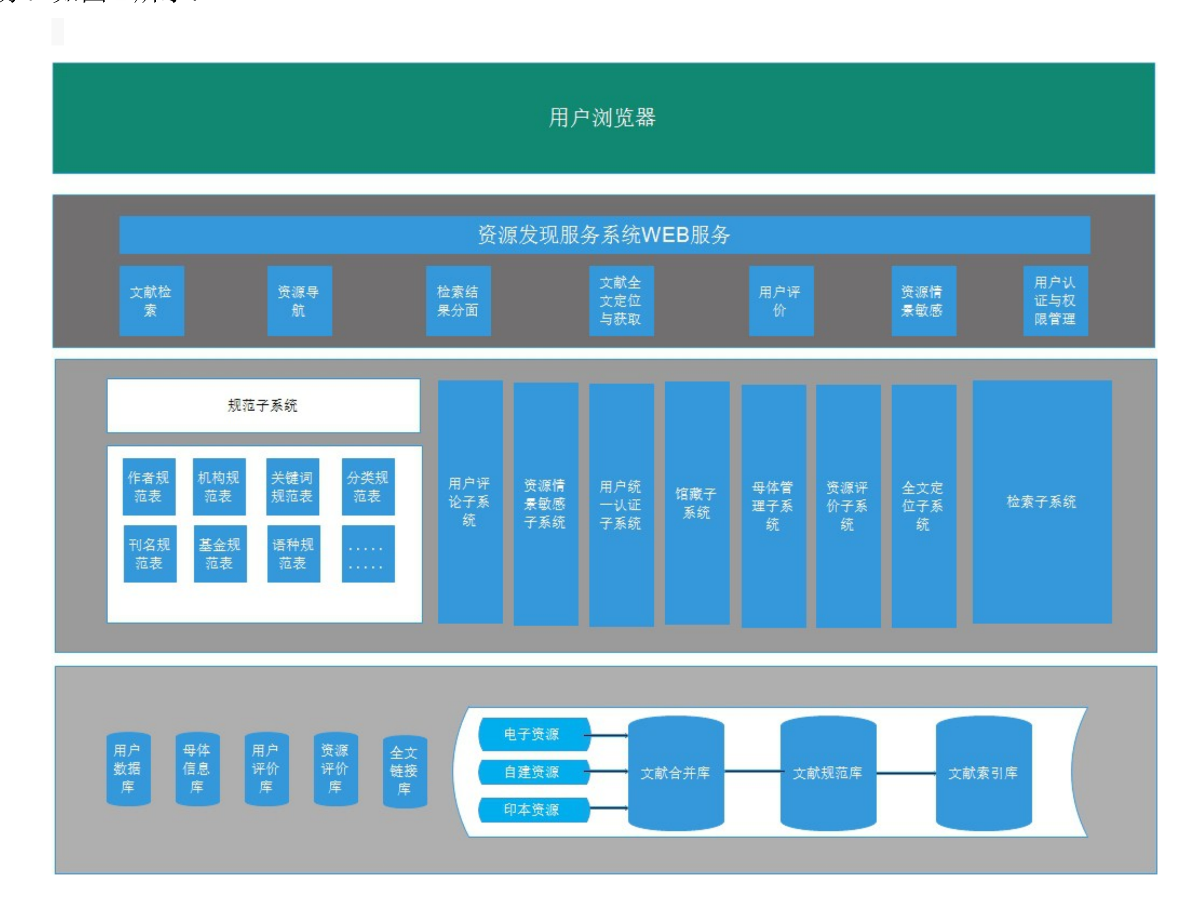


图 1：文献资源发现服务系统的技术框架

数据层

数据层包括支撑系统运行的若干底层数据库，包括用户数据库、母体信息库、用户评价库、资源评价库、全文链接库和文献数据库。这些数据库分别支撑着业务层的相关业务逻辑模块。其中文献数据库是系统中最重要的数据库。下面详细介绍一下文献数据库的加工流程。

1. **电子资源导入到合并库**

   电子资源提供方可能提供各种格式异构数据格式。为了能够正常导入到合并库中，我们需要对各种异构数据格式进行分析，建立异构数据格式对合并库的抽取和映射关系。可以用专用工具直接将易购数据导入到合并库中，如果映射关系比较复杂，也可以开发专门的工具软件来导入特定的电子资源。电子资源另一种进入合并库的方式是收割。比如可以开发专用收割程序，定时将特定网站上的电子资源按更新日期收割到合并库

中。

如果电子资源的规范性、准确度较差，可能还要编写数据清洗程序对数据进行筛选，剔除不符合要求的数据。必要时要进行人工处理。

为了使系统的数据收割/导入工作运行得更平稳和有效率,应该建立比较规范的数据导入流程。一种比较好的流程可以采取以下方式：定义统一的导入元数据规范，规范中规定了统一的数据结构和各个字段项内容的规范。各电子资源提供方将数据转换为符合此规范的格式，一般为XML文件。系统提供统一的导入接口，由提供方自行导入，也可以指定从特定的源进行收割。

2. **数据的合并、查重**

由于合并库有各个元数据提供者提供，这不可避免的导致库内存在重复数据。为了提高客户的体验，合并、去重工作十分重要。

可以以若干组关键标识字段座位判断元数据重复的标准。如 DOI 字段可以单独作为一个去重标准；标题、作者、年份、页码可以作为一个去重标准。按照这些组标准确定不同元数据之间是否存在重复关系。

确定若干条元数据之间的重复关系后，下一步工作就是将多条元数据合并为一条。一个比较简单的策略是按照元数据提供方的规范程度，直接选取规范程度较高的提供方的元数据作为合并后的元数据。另一种更为完善的方法则以规范程度较高的提供方的元数据作为基准，逐个字段比较重复元数据的规范程度。将规范度高的记录的字段内容填入基准元数据记录的相应字段，必要时进行人工处理。这种方式可以提高合并后元数据的质量，但是降低了效率，并提高了系统复杂度。

3. **数据内容规范**

系统中有专门负责数据内容规范的子系统，其中包含了作者姓名、机构、关键词、基金、学科分类、刊名、语种、出版社等规范表。这些表种包含了字段规范内容和非规范内容的对应关系。当系统中导入新数据时，数据规范子系统就可以根据这些表对新数据的不规范内容进行规范。数据规范表应该不断更新，保证数据的准确和有效性。

4. **建立索引、提供索引服务**

为了让用户快速检索到自己想要的文献元数据，必须对合并、规范后的数据建立索引。根据用户的检索需求，分析出应该建立索引的元数据字段。一般通用的单一检索框需要对所有检索字段建立全文索引；如果系统提供高级检索模式，那么为了满足对某字段的精确检索需求，需要对整个字段内容建立非全文索引。索引文件建立后可以通过全文检索引擎对外提供检索服务，全文检索引擎按照功能大致可以分为检索模块和存储模块。当用户提交检索条件时，检索条件种的检索内容将会根据语言进行适当的分词后，提交给全文检索引擎中搜索模块进行检索，搜索模块会通过全文检索引擎中的存储模块直接访问索引文件进行查找，再将搜索到的内容返回给用户。

应用层

根据功能划分，应用层根据分为若干子系统：

1. 规范子系统
   其中包含多种规范表，负责对文献库中若干字段内容进行规范。
2. 用户评论子系统
   负责用户对文献的评论和打分
3. 用户统一认证子系统

负责系统用户的认证和权限管理，如系统内某些特定资源只对部分用户开放。

4. 资源情景敏感子系统

   负责根据用户所在环境是否在馆内，决定资源的使用权限。某些资源只能在图书馆内环境下使用。还可以根据用户所在的为止提供文献所在附近图书馆的馆藏情况，以便到馆借阅。

5. 馆藏子系统

   记录了各个电子资源提供方的馆藏情况，并在母体层面进行了规范和去重。当用户找到一篇文献资源时，如果资源有多个馆藏，会同时列出供用户选择。

6. 母体管理子系统

   对所有电子资源的母体信息进行编目,并对到馆信息进行登到。

7. 资源评价子系统

   与用户评价子系统不同，资源评价子系统是根据一系列科学评价的客观指标，利用一系列工具或插件，对系统里的母体或文献进行评价。目前比较著名的分析工具有 EBSCO 公司的 PLUMX。

8. 全文定位子系统

   使用静态或动态定位方法，定位用户请求文献全文的网络地址。如果一篇文献有多个馆藏，可以让用户选择请求特定的馆藏全文。

9. 检索子系统

   提供文献检索服务。这是系统里最基础、最重要的服务，资源发现服务的核心功能。

表现层

表现层即 WEB 服务，由若干前台页面组成，根据不同的功能划分可以分为若干组：

1. 文献检索

   包括单一输入框检索，和适用于更专业人员使用的多条件检索和高级表达式检索页面。另外还包括检索结果页面。

2. 资源导航

   可以根据刊名、会议名、分类号等字段分别对不同文献类型文献进行导航，帮助用户快速浏览自己想要的文献。

3. 检索结果分面

   根据多个指标对文献检索结果进行分面显示，如出版年、分类、文献类型、刊名等。

4. 文献全文定位与获取

   提供一组页面供读者获取自己想要的全文链接，如果找不到全文链接，应该有页面提供原文传递服务。

5. 用户评价

   用户可以在相关页面对某篇文献发表评论或评分。

6. 资源情景敏感

   在页面根据用户所在位置推荐附近的馆藏资源，以便到馆借阅相关文献。

7. 用户认证与权限管理

   包括用户登录、注册与个人信息管理页面，某些特定资源的显示与获取也与当前用户的权限有关。

3、 资源发现服务系统的技术要点

在资源发现服务系统的建设过程中，需要一些关键技术的支持，下面分别介绍一些主要的关键技术。

1. 全文检索技术

在上面对资源发现服务系统的分析中，我们可以看到用户使用最频繁的模块仍然是文献检索功能，这可以看作是决定系统成败的核心功能。所以选择一种能够准确、快速在海量文献检索出用户所需要的文献的全文检索引擎尤其重要。 从整个用户群体的需求来看全文检索引擎要有如下几个特点：

1/ 检索响应速度要快捷。

用户的体验好坏直接取决于返回检索结果的时间长短，要提高用户体验，返回时间越短越好。在目前的大数据时代，文献检索系统种的数据量也是非常庞大的。在海量数据中检索，保证响应速度尤其重要。

2/ 要有优秀的中文分词算法。

分词是全文检索的基础工作，而中文的词之间没有分隔符，这与英语等西方语言有明显的区别。中文只是字、句和段能通过明显的分界符来简单划分，唯独词没有一个形式上的分界符，虽然英文也同样存在短语的划分问题，但是在词这一层上，中文比之英文要复杂和困难的多[3]由于中文分词技术是自然语言检索里的基础性技术，目前相关研究比较多。在文[4]中，将中文分词算法分为三大类：基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法。介绍了中文分词的发展现状及在搜索引擎中的运用。

3/叙词表在中分分词中的利用

与普通互联网检索引擎不同，科技文献中含有大量的专业词汇，使用通用中文分词算法并不能取得很好的效果。叙词表又称为主题词表，它是一种语义词典，由术语及术语之间的各种关系组成，能反映某学科领域的语义相关概念。[5]我们可以把虚词表中的术语表作为一种专业词库，加入到全文分词的自定义词库中。这将明显提高分词的准确性。文[6]中提出了将医学叙词表 MeSH 词汇加入到通用分词表中进行分词，并利用 MeSH 词汇结合词长、词语所在位置加权实现医学新闻网页的关键词自动提取策略。

2. 云计算技术

由于资源发现服务系统中的文献数据量特别巨大，其最终形成的索引库也是非常庞大的。靠单机往往难以承担重任。此时的索引方案是将整个文档集合划分为若干子集，建立分布式集群，即每台机器维护整个索引的一部分，由多台机器共同完成索引的建立和对检索的响应。[7]当用户发出检索请求时，一台检索分发服务器将请求分发给多台检索服务器。每台检索服务器完成检索后将结果返回给分发服务器，再经过合并、排序后返回给用户结果。目前主流的全文检索引擎都支持这种并行计算扩展。随着资源发现服务系统用户数量增多、文献数据量不断增大，系统所需要的服务器资源也不断增加。我们可以把系统种的检索模块部署在云计算平台，根据系统访问量和数据量的变化动态调整用于并行检索计算的服务器数量。这样较传统的数据中心机房有降低运营成本、动态可扩展、简化维护等优势。

3. 全文定位获取机制

用户访问资源发现服务系统的最终目的是获取到自己需要的文献全文，做好这最后一步需求是系统成败的关键之一。可以采取多种机制定位和获取全文。

1/对于本馆的电子文献资源

这是最容易获取全文的一种资源，因为资源在本馆存放，直接根据文献全文地址定位获取即可。

2/有 DOI 的文献元数据

DOI 是"Digital Object Identifier"的简写，用来标识在数字环境中的内容对象。通过文献的

DOI 可以获取文献的全文数据。在浏览器地址栏输入 http://dx.doi.org/，在"Resolve A DOI Name"的提示框内输入文献 DOI, 点击"Go"按钮，DOI 系统就会自动链接到该文献的 url,并显示相应的页面。如果访问者购买了该文献所在数据库的资源使用权，则可以直接下载全文；如果没有，可能需要单篇购买。

3/由电子资源提供方提供的文献 ID 或者文献全文 URL

可以直接使用文献电子资源提供方提高的 URL 地址获取全文；也可以通过提供方的文献 ID， 按照一定的格式拼接出全文的 URL 地址。使用此方法的缺点是全文地址的 URL 可能会发生变化，导致获取失败。为了避免这种情况，需要系统定期更新那些失效的全文链接内容。

4/动态全文获取

通过程序代码根据文献数据的关键字段（题名、作者、出版时间等）在资源提供方系统内进行检索，将最符合检索条件的文献数据的URL地址返回用户。这是一种动态全文地址解析过程。这个过程可以是离线完成的，即系统定时在后台进行动态解析并更新；也可以在用户发出请求后实时进行解析。这种方法的缺点是准确率的问题，由于检索并不总是能返回正确的文献元数据，所以可能会找到错误的结果。

文献资源发现服务系统应该综合采用上面各种机制来帮助用户定位和获取全文，在此过程中尽可能提供最好的用户体验。

4. 检索结果排名算法设计

用户在海量数据中进行检索，得到的也可能是一个很大的检索结果集。如何将用户真正需要的文献排在显示结果中的前面，也是决定着系统成败的一个关键点。排序算法会算出每篇文献的排名指数，应该从两个方面设计。

首先是计算文献与检索条件之间的相关度。如果用户检索文献题名与检索结果中文献的题名完全一致，那么相关度为最高。如果不完全匹配，要将检索条件分词后分别计算每个词在文献元数据字段中的词频。这些字段应该包括文献题名、主题词、关键词、摘要等字段。再按照一定的权重计算出一个总的权重，代表了文献与检索条件的相关程度。

其次是考虑一些文献或者文献母体的指标，如出版日期、文献类型、文献被引用次数、母体文献被引用次数、是否为图书馆自由资源、文献在系统内被浏览和被请求全文次数等。还可以参考一些学术性/同行评论，并结合系统自身的读者评价功能。

最后的排序指数是按照上面的各种指标按照一定的加权系数相加得出，加权系数可能要根据系统运行情况或者用户意见进行调整。

4、 总结与展望

图书馆检索系统总是随着时代而发展。由传统系统到联邦检索，又从联邦检索发展到资源发现系统。当前的文献资源发现服务系统适应了互联网时代，充分考虑了用户的检索习惯，为读者请求所需全文提供最大的便利。未来的资源发现系统也将以用户需求与体验为第一要旨。当前的检索系统本质上仍然是基于检索词进行检索，随着搜索引擎技术的发展，未来基于语义的检索可能会应用到资源发现系统中。随着移动互联网的普及，文献资源发现系统应该移植到移动设备上，使得用户可以随时访问系统。用户应该可以更多进行个性化定制，如用户如果对检索结果的排序不满意，可以自定义排序算法或者调整排序指标的加权系统等。

参考文献

[1]马骅 .国外主要联邦检索系统的兴起、现状及发展趋势[J].图书馆建设,2009,(3):1-5.
[2]陈家翠. 联邦检索机制及其存在的问题[J]. 图书情报工作,2006,50(6): 87-89,103.

[3]王星. 国家工程技术数字图书馆技术体系[J]. 数字图书馆论坛,2013,(10):14-19.
[4]何莘,王琬芫. 自然语言检索中的中文分词技术研究进展及应用[J].情报科学,2008,26(5): 787-797.
[5]李景,钱平.叙词表与本体的区别与联系[J].中国图书馆学报,2004,30(1): 36-39.
[6]何晓阳,张精理,丁婷.医学新闻关键词自动提取策略[J].中华医学图书情报杂志,2014,(4): 13-17.
[7] 王灏,张正锋,冯巍.图情资源发现系统的研究与实现[J]. 数字图书馆论坛,2013,(6): 51-56.

Discussion on the Realization of the Literature Resource Discovery Service System

Liu MinJian　　Wang Xing
(Institute of Scientific and Technical Information of China， Beijing 100038)

Abstract: This paper briefly introduces the basic concepts of literature resource discovery service system, and introduce the technology framework of the system in detail according to the basic principle,and one of the key technologies used are briefly described.

Keywords: Resource Discovery, Digital Library, Technology Framework, Key Technology